GEOPERTURB: Probing Structural Vulnerability in Multimodal Geometry Reasoning

Anonymous ACL submission

Abstract

Large Multimodal Models have demonstrated remarkable advancements in geometric reasoning, yet it remains fundamentally unclear whether their performance stems from robust abstract deduction or fragile template memorization. In this paper, we challenge the limits of LMMs by presenting the first systematic investigation into their structural vulnerability. Rather than relying on low-level, pixel-level corruptions, we introduce GeoPerturb, an integrated diagnostic framework that treats programmatic diagrammatic reconstruction as a controlled representation intervention. Under this framework, we curate a highly refined, 531-problem benchmark, wherein each geometric problem is systematically modulated across four structural perturbation families: Orientability, Referentiality, Appearance, and Redundancy. Extensive evaluations on frontier LMMs uncover severe and pervasive logical blind-spots. We show that while modern models often excel at canonical layouts, many of them still suffer substantial and family-dependent degradation under semantically invariant transformations. These empirical discoveries reveal a deep vision-logic gap, providing compelling evidence that current LMMs can still rely on spurious visual heuristics and public-template shortcuts rather than robust mathematical deduction.

1 Introduction

The rapid convergence of computer vision and natural language processing has fueled the rise of Large Multimodal Models (LMMs), positioning them as core engines for advanced intellectual tasks (Yan et al., 2025; Zhong et al., 2026). Among these, automated mathematical reasoning, specifically geometric problem solving, serves as a crucial frontier for evaluating an agent's high-order cognitive capabilities (Xu et al., 2024; Chen et al., 2025; Li et al., 2026; Cho et al., 2026; Ping et al., 2026). Geometric deduction requires a sophisticated interplay between perceptual parsing and logical execution: a model must not only recognize visual primitives such as lines, angles, and curves from a diagram but also map them precisely to the abstract mathematical constraints articulated in the textual prompt. Driven by massive-scale pre-training and alignment, current state-of-the-art LMMs (Zhan et al., 2025; Yang et al., 2026a) have achieved unprecedented scores on static geometry benchmarks, sparking optimism that these systems have begun to master genuine mathematical deduction (Yang et al., 2026b). Despite these performance gains, however, it remains unclear whether such success stems from deep deductive reasoning or merely from the memorization of specific visual configurations prevalent in static benchmarks.

Recent studies in language domain have demonstrated that LLMs often rely on superficial text patterns rather than consistent logical deduction (Wang et al., 2021; Zhu et al., 2024), and this concern is even more pronounced in the multimodal context. While research has begun to extend such robustness evaluations to multimodal settings, existing efforts primarily focus on perceptual-level perturbations, such as adversarial pixel noise (Cui et al., 2024) or generic image corruptions (Qiu et al., 2022). These pixel-centric evaluations fail to account for the unique structural invariants inherent in geometric reasoning. Unlike natural scenes, geometric diagrams are visual embodiments of rigid logical constraints, where the mathematical essence must remain invariant under various structural transformations. Current benchmarks, being predominantly static, fail to probe whether LMMs can maintain reasoning consistency across structurally equivalent variants.

To bridge this critical gap, we present GeoPerturb, a programmatic framework designed to systematically probe the structural vulnerability of modern LMMs. Instead of treating diagrams as im-

mutable pixel arrays, we introduce a paradigm shift by leveraging TikZ-mediated reconstruction to decode unstructured images into verifiable, parametric code. This formal intermediate representation allows us to decouple the underlying mathematical logic from its superficial visual rendering. Upon this canonical baseline, GeoPerturb automates precise, logic-preserving transformations across four distinct structural axes: Orientability (spatial rotation), Referentiality (symbolic permutation), Appearance (visual stylization), and Redundancy (irrelevant injection).

Our extensive empirical evaluation reveals a profound "vision-logic" gap in frontier models: while a model may solve a canonical problem flawlessly, its reasoning can degrade substantially under these isomorphic shifts. This structural fragility underscores that current LMMs heavily exploit spurious visual correlations rather than internalizing invariant mathematical truths, providing a rigorous diagnostic lens for future development in verifiable multimodal AI.

In summary, our main contributions are:

- **Identifying Vulnerability:** We present the first systematic investigation into the structural vulnerability of LMMs, shifting the focus from pixel-level noise to invariant logical consistency in multimodal geometry reasoning.
- **Constructing GeoPerturb:** We introduce GeoPerturb, a diagnostic framework for evaluating the structural robustness of LMMs. The GeoPerturb dataset comprises 531 carefully curated geometry problems, each containing a canonical reconstruction and four structural perturbation variants.
- **Uncovering Blind-spots:** We show that structural robustness is highly model- and family-dependent: high canonical accuracy can coexist with large joint-consistency collapse and conditional grounding failures, revealing where spurious layout shortcuts persist.

2 Related Work

Reasoning Benchmarking. Recent advancements in LMMs have significantly expanded the horizon of automated reasoning, particularly in complex domains such as geometry (Xu et al., 2024; Zhao et al., 2025b; Fu et al., 2025; Shi et al., 2025; Zhang et al., 2025; Wu and Sun, 2025). Existing benchmarks including GeoQA (Chen et al., 2021) and Geometry3K (Lu et al., 2021), have been instrumental in assessing the ability of models to translate visual information into mathematical proofs. However, these benchmarks predominantly rely on static diagrammatic corpora, where models often achieve high performance through dataset-specific heuristic learning rather than generalized reasoning (Zhu et al., 2026). The inherent limitation of these fixed-template datasets is their inability to isolate logical understanding from incidental visual memorization. Consequently, there is an urgent need for diagnostic benchmarks that can probe whether models exhibit invariant reasoning across semantically equivalent but structurally transformed geometric representations.

Robustness Probing. The robustness of vision-language models has been extensively studied through various perturbation lenses, including adversarial attacks (Goodfellow et al., 2014; Qi et al., 2024), image corruptions (Hendrycks and Dietterich, 2019; Zhu et al., 2026), and synthetic noise injection (Fu et al., 2024). These studies have provided critical insights into how pixel-level perturbations can lead to catastrophic failures in downstream tasks. Despite these contributions, existing robustness probes remain heavily tethered to low-level signal distortion. Recent efforts have begun to touch upon structural invariance in vision tasks (Zhao et al., 2025a), yet a rigorous exploration of this phenomenon within the context of geometric reasoning, where structure is paramount, remains sparse. Our work bridges this gap by introducing structural diagnostic probing, which evaluates models not against random noise, but against controlled transformations of the underlying logical skeleton.

Programmatic Synthesis. The integration of programmatic data generation, particularly using languages like TikZ, has emerged as a powerful paradigm for augmenting training data and enhancing compositional reasoning (Belouadi et al., 2024a,b; Wei et al., 2025; Greisinger and Eger, 2026). By leveraging code as a medium to define geometric properties, researchers have successfully generated synthetic datasets that bypass the limitations of human-annotated samples (Saito et al., 2025; Sun et al., 2026). While these programmatic methods have proven effective in scaling up training regimes, their utility as diagnostic instruments—specifically for decoupling representational capacity from structural reasoning—remains largely untapped. Our framework, GeoPerturb, advances this paradigm by utilizing programmatic

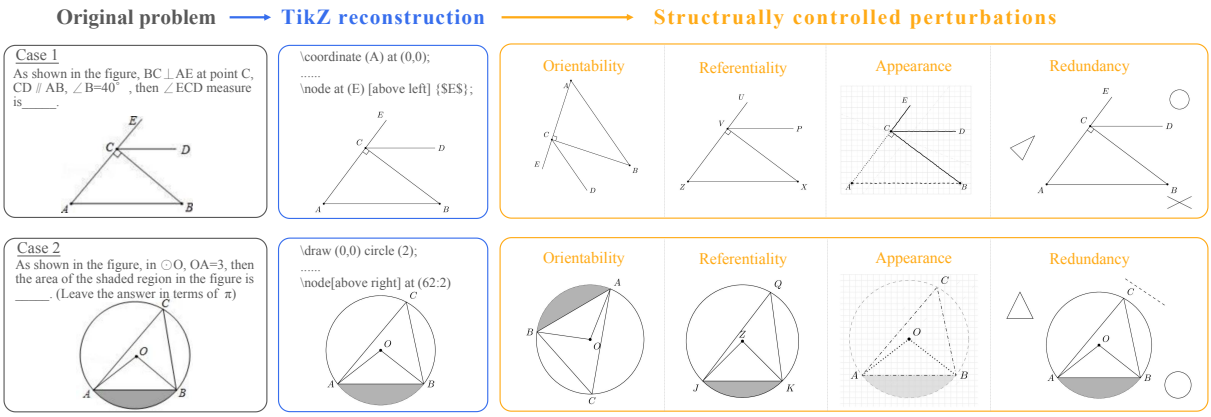

Figure 1: The GEOPERTURB framework illustrating four structrualy controlled perturbations.

reconstruction as a systematic intervention, providing a precise mechanism to stress-test the structural stability of modern multimodal architectures.

3 The GEOPERTURB Framework

3.1 Problem Formulation

Let $P = (I, T, A)$ denote a canonical geometric problem, where I represents the vector or raster geometry diagram, T signifies the textual description and question prompt, and A is the ground-truth answer. Traditional evaluation paradigms implicitly assume that an LMM $\mathcal{M}$ possesses geometric reasoning capabilities if it satisfies $\mathcal{M}(I, T) = A$.

However, this mapping conflates genuine mathematical deduction with spatial heuristic memorization. To decouple these factors, we formalize the evaluation as a structural stress-test. We leverage a programmatic synthesis agent $\mathcal{G}$ that maps a textual specification into an editable vector format (e.g., TikZ) and then redering to a canonical reconstruction. For each problem, we generate a canonical reconstruction $I_{\text{can}} = \mathcal{R}(\mathcal{G}(I, T))$ and a set of structural variants via a transformation tensor:

$$I_{\text{per}} = \mathcal{T}(I_{\text{can}}, \theta),$$

where $\mathcal{T}$ represents a semantic-preserving operator parameterized by transformation configurations θ. A model is defined as structurally robust if and only if its predictive alignment holds invariant under all valid perturbations:

$$\mathbb{P}\big(\mathcal{M}(I_{\text{per}}, T) = A\big) = \mathbb{P}\big(\mathcal{M}(I_{\text{can}}, T) = A\big), \forall \mathcal{T}.$$

3.2 Perturbation Taxonomies

We construct four distinct families of semantic-preserving transformations, as visually conceptualized in Figure 1. While these geometric modifications strictly maintain *structural invariance*, preserving the underlying logical and mathematical truth of the problem—they deliberately expose the model's *structural vulnerability* by disrupting superficial visual heuristics. Benefiting from our programmatic TikZ backbone, these interventions are executed with high precision at the compilation level, decoupling spatial, symbolic, stylistic, and redundant components that are inextricably tangled in standard raster images.

- **Orientability** (*Spatial Rotation*): This family enforces orientation invariance by applying unconstrained, randomized rigid transformations to the geometric skeleton, via discrete rotations. Crucially, leveraging the semantic decoupling capability of TikZ, *these operations rotate the geometric shapes exclusively while keeping the textual vertex labels fixed in their upright canonical orientations*. This isolates the spatial configuration from alphanumeric distortion, testing whether LMMs rely on fixed-orientation layout templates.
- **Referentiality** (*Symbolic Permutation*): This dimension evaluates grounding invariance under arbitrary isomorphic shifts. We systematically permute the literal alphabet labels assigned to geometric vertices (e.g., mapped from a canonical $\triangle ABC$ to an unconstrained, randomized identity $\triangle XYZ$). The mapping enforces a strict bijective (one-to-one) correspondence between the new randomized character tokens and their original geometric coordinates, keeping the logical deductive chain intact while completely breaking any reliance on high-frequency visual-textual token co-occurrences memorized from web-scale data.

- **Appearance** (*Visual Stylization*): Inspired by Gestalt psychology (Koffka, 2013), this family introduces stylistic invariance probes by altering the visual surface rendering without shifting a single pixel of the topological skeleton. We inject heterogeneous artistic modifications, including rendering uniform background coordinate grids, altering stroke color semantics, converting continuous solid lines to dashed fragments, and transmuting precise vector renders into organic hand-drawn sketch styles. This stresses the model's ability to extract abstract shapes independent of low-level visual textures.
- **Redundancy** (*Irrelevant Injection*): This taxonomy evaluates information-sufficiency invariance amidst visual clutter. Without altering the original region of interest or modifying the core logical paths required for the solution, we inject non-essential auxiliary geometric primitives, such as redundant tangent lines, arbitrary bounding boxes, unlinked circles, and isolated witness points, into mathematically unconstrained vacant spaces of the canvas. This directly diagnoses whether structural redundancy imposes an asymmetric cognitive overload on the model's visual attention mechanism.

3.3 Data Construction

To construct our dataset repository, we meticulously source geometric problems from both established benchmarks and independent human collections, yielding a total of 531 matched problem suites. Each suite functions as a tightly coupled visual-logical cluster containing the original source raster, our compiled canonical TikZ template, and four programmatic structural variants. To systematically diagnose whether models rely on layout shortcuts or suffer from dataset contamination, we stratify our data into three explicit splits based on their assumed exposure in public web domains. Specifically, the **Train Split** ($N = 347$) is sourced directly from the training subsets of prominent geometry datasets, namely GeoQA (Chen et al., 2021) and Geometry3K (Lu et al., 2021), serving as an empirical control group to evaluate whether models exhibit inflated performance on highly accessible, web-scraped public distributions. Concurrently, the **Test Split** ($N = 96$) is drawn from the corresponding evaluation subsets of GeoQA and Geometry3K; analyzing the discrepancy in performance drops between the Train and Test splits under identical structural perturbations allows us to audit the baseline generalization boundaries within a shared dataset distribution. Complementing these open-source subsets, the **Pure Split** ($N = 88$) is curated as a novel evaluation anchor to mitigate the pervasive risk of data contamination in closed-source frontier LMMs. This split comprises advanced, hand-collected geometry problems. While absolute immunity to pre-training data ingestion cannot be fully guaranteed for proprietary, web-scale foundation models, the `Pure` split serves as a highly leakage-mitigated, out-of-distribution pressure-test to differentiate genuine mathematical reasoning from mere template retrieval.

To ensure absolute visual and mathematical precision before conducting large-scale model probing, all 531 suites undergo a rigorous two-stage human audit. First, a geometric fidelity inspection is performed on every canonical TikZ reconstruction to verify that topological relations and vertex label placements are perfectly preserved. Rather than discarding compilations that failed initial automated scripts, we manually correct 73 complex diagrams to eliminate selection bias. Second, all perturbed variants are verified to ensure that the semantic core remains invariant and that the gold standard answers are correctly maintained.

4 Experiments

Evaluated Models. To ensure a comprehensive and representative diagnosis of contemporary vision-language capabilities, we construct an evaluation suite spanning both industry-leading proprietary foundations and state-of-the-art open-source architectures. Our evaluation matrix includes:

- **Proprietary Models:** We benchmark a diverse cohort of commercial closed-source models, evaluating the frontier capabilities of `GPT-5.4-High` (OpenAI, 2026), `GPT-4o-mini` (OpenAI, 2024), `Claude-4.6-Opus` (Anthropic, 2026a), `Claude-4.6-Sonnet` (Anthropic, 2026b), `Gemini-3.1-Pro-Preview` (Google DeepMind, 2026), and `Qwen3-VL-Plus` (Bai et al., 2025). This tier represents the current ceiling of multimodal abstract deduction and commercial spatial reasoning.
- **Open-Source Models:** To facilitate reproducibility and analyze accessible structural representations, we extend our evaluation to prominent open-source vision-language models across varying parameter scales,

Table 1: **Macro Benchmarking Results.** We evaluate geometric proficiency across original scrapes (`Original`), precise vector reconstructions (`Canonical`), and semantic-preserving transformations (`Perturbed`). Crucially, `Perturbed` reports the *joint accuracy*, where an instance is marked correct only if it passes all four perturbations. All accuracies are rounded to one decimal; delta annotations are computed before rounding.

Model	Original			Canonical			Perturbed		
	train	test	pure	train	test	pure	train	test	pure
Proprietary Models									
GPT-5.4-High	90.2	88.6	65.5	88.1	87.5	67.4	71.5 (↓16.6)	60.0 (↓27.5)	51.2 (↓16.2)
Gemini-3.1-Pro	95.7	98.0	76.1	93.0	94.8	69.3	80.9 (↓12.1)	75.2 (↓19.6)	50.0 (↓19.3)
Qwen3-VL-Plus	87.3	88.3	66.7	87.5	89.6	70.1	67.3 (↓20.2)	70.3 (↓19.3)	42.4 (↓27.7)
Claude-4.6-Opus	89.6	84.4	53.4	89.6	89.6	59.1	88.1 (↓1.5)	85.8 (↓3.8)	58.0 (↓1.1)
Claude-4.6-Sonnet	83.9	81.2	54.5	83.6	82.3	63.6	81.3 (↓2.3)	85.4 (↑3.1)	57.7 (↓5.9)
GPT-4o-mini	41.0	38.6	29.5	41.2	40.6	29.5	20.6 (↓20.6)	17.8 (↓22.8)	14.0 (↓15.5)
Open-source Models									
Qwen3-VL-8B	61.8	63.5	37.3	55.9	64.6	32.8	53.7 (↓2.2)	60.4 (↓4.2)	33.1 (↑0.3)
InternVL-3.5-38B	54.0	61.5	39.8	51.3	59.4	47.7	48.2 (↓3.1)	57.8 (↓1.6)	42.6 (↓5.1)
LLaVA-OneVision-8B	36.8	41.7	23.9	36.0	42.7	25.0	31.8 (↓4.2)	42.9 (↑0.2)	25.0 (↓0.0)
MathCoder-VL-8B	40.8	45.8	28.4	36.3	41.7	26.1	32.3 (↓4.0)	37.3 (↓4.4)	27.3 (↑1.2)
G-LLaVA-13B	16.7	30.2	25.0	17.0	32.3	23.9	16.7 (↓0.3)	30.4 (↓1.9)	21.9 (↓2.0)

including the `Qwen3-VL-8B` (Bai et al., 2025), the `InternVL-3.5-38B` (Wang et al., 2025b), and the unified multimodal baseline `LLaVA-OneVision-8B` (An et al., 2025). Furthermore, to investigate the impact of domain-specific mathematical and geometric pre-training, we incorporate specialized open-source baselines: `MathCoder-VL-8B` (Wang et al., 2025a) and the geometry-focused `G-LLaVA-13B` (Gao et al., 2025).

Prompting Strategy and Metrics. To elicit pure zero-shot spatial reasoning and ensure deterministic scoring, all models are queried using a standardized, structure-forcing prompt template. Models are instructed to output their final answer within a designated JSON block containing a single numerical or option token, matching our compiled gold-standard answers.

Crucially, to probe model robustness under visual drift, we report the performance on the perturbed split using a rigorous metric that an instance in the Perturbed split is marked as correct if and only if the model simultaneously solves all four distinct programmatic layout transformations (Orientability, Referentiality, Appearance, and Redundancy) applied to that specific problem.

4.1 Macro Analysis: Global Vulnerability

The primary quantitative performance across our three evaluation tiers is reported in Table 1.

Canonical Layout: Volatile Equivalence. A comparative audit between the web-sourced `Original` rasters and our reconstructed `Canonical` configurations (columns 2–4 vs. 5–7 in Table 1) reveals a state of high-variance, dynamic equivalence. Because our `Canonical` split utilizes programmatic TikZ syntax to render clean imagery, it serves as a strict vector-graphic baseline for subsequent downstream perturbations. When shifting to this TikZ-rendered paradigm, individual architectures exhibit non-monotonic localized fluctuations rather than a uniform trend.

For instance, on the leak-mitigated `Pure` split, Claude-4.6-Sonnet experiences a substantial accuracy surge (54.5% → 63.6%). This improvement indicates that clean vector rendering successfully eliminates low-level raster noise, compression artifacts, and pixel blurring present in web scraped images, thereby facilitating cleaner geometric edge detection for the vision encoder.

Conversely, Gemini-3.1-Pro undergoes a notable performance regression (76.1% → 69.3%). This decay implies that certain closed-source models suffer from severe over-fitting to micro-rendering artifacts specific to public web benchmarks. Geometric diagrams scraped from the web carry distinct low-level footprints, including specific font rendering engine outputs for text labels, sub-pixel anti-aliasing profiles along line paths, and standardized layout padding. Overall these positive and negative architectural shifts cross-cancel each other across the entire cohort, the benchmark maintains macro aggregate parity.

Perturbed Layout: Heterogeneous Structural Sensitivity. In contrast to the localized fluctuations under canonical baselines, transitioning to the `Perturbed` configuration exposes substantial but model-dependent structural sensitivity across

Table 2: **Micro Benchmarking Results.** We decompose geometric proficiency across four fine-grained perturbation families, aggregated as a dataset average across the evaluation spectrum. Crucially, these marginal accuracies isolate decoupled sub-competencies, Orientability, Referentiality, Appearance, and Redundancy.The reported averages are computed over matched valid instances rather than as the unweighted mean of the train, test, and pure columns.

Model	Canonical	Orientability	Referentiality	Appearance	Redundancy
Proprietary Models					
GPT-5.4-High	84.6	80.3 (↓4.3)	83.8 (↓0.8)	82.5 (↓2.1)	86.4 (↑1.8)
Gemini-3.1-Pro	89.4	84.2 (↓5.3)	89.2 (↓0.2)	87.7 (↓1.7)	90.4 (↑1.0)
Qwen3-VL-Plus	84.9	78.5 (↓6.4)	84.7 (↓0.2)	83.1 (↓1.8)	85.0 (↑0.1)
Claude-4.6-Opus	84.5	77.8 (↓6.8)	84.0 (↓0.5)	84.9 (↑0.3)	83.6 (↓1.0)
Claude-4.6-Sonnet	80.0	73.6 (↓6.4)	77.8 (↓2.3)	80.7 (↑0.6)	80.0 (↓0.1)
GPT-4o-mini	39.1	38.1 (↓1.1)	38.3 (↓0.8)	38.3 (↓0.8)	38.8 (↓0.3)
Open-source Models					
Qwen3-VL-8B	54.5	50.4 (↓4.1)	52.3 (↓2.3)	49.9 (↓4.6)	55.3 (↑0.8)
InternVL-3.5-38B	52.2	46.8 (↓5.4)	47.7 (↓4.4)	50.1 (↓2.1)	50.0 (↓2.2)
LLaVA-OneVision-8B	35.4	31.4 (↓4.0)	30.1 (↓5.3)	33.6 (↓1.8)	34.2 (↓1.2)
MathCoder-VL-8B	35.6	33.3 (↓2.3)	33.2 (↓2.4)	28.5 (↓7.1)	33.8 (↓1.8)
G-LLaVA-13B	20.9	21.8 (↑0.9)	14.3 (↓6.6)	20.2 (↓0.7)	22.1 (↑1.2)

the three data allocations. Because both the `Canonical` and `Perturbed` configurations share identical programmatic TikZ rendering pipelines, this comparison isolates layout structure as the sole experimental variable. For several proprietary flagships, the absolute degradation peaks dramatically on the `Test` split, as exemplified by GPT-5.4-High dropping by an absolute 27.5% (87.5% $\rightarrow$ 60.0%) and Qwen3-VL-Plus regressing by 19.3% (89.6% $\rightarrow$ 70.3%). At the same time, the Claude results reveal that this vulnerability may not be uniform: Claude-4.6-Opus drops by only 3.8 points on the `Test` split, and Claude-4.6-Sonnet even improves slightly under the joint perturbation metric.

For the open-source cohort, the updated run shows much milder joint-perturbation drops than the earlier sweep. On the `Test` split, Qwen3-VL-8B declines from 64.6% to 60.4%, InternVL-3.5-38B from 59.4% to 57.8%, while LLaVA-OneVision-8B remains nearly unchanged. These results indicate that the joint metric is not uniformly catastrophic, even though the family-level analysis below still reveals persistent structural bottlenecks.

4.2 Micro Analysis: Dimensional Bottlenecks

To establish a more granular understanding, we systematically deconstruct model performance across four isolated perturbation dimensions: Orientability, Referentiality, Appearance, and Redundancy. To filter out localized variance and capture the overarching structural sensitivity of each model, our analysis reports the matched-instance aggregate accuracy across the available evaluation splits. The comprehensive comparisons are presented in Table 2.

Orientability: Spatial Anchors Are Broken. Spatial orientation remains the most consistent destructor of layout shortcuts across the model spectrum. When diagrams undergo programmatic rigid rotations, proprietary models exhibit profound performance drops relative to their `Canonical` baselines, visible in Gemini-3.1-Pro (89.4% $\rightarrow$ 84.2%), Qwen3-VL-Plus (84.9% $\rightarrow$ 78.5%), Claude-4.6-Opus (84.5% $\rightarrow$ 77.8%), and Claude-4.6-Sonnet (80.0% $\rightarrow$ 73.6%).

Crucially, because our TikZ setup rotates the geometric pipelines exclusively while keeping the alphanumeric text labels fixed in their upright canonical orientations, this degradation cannot be attributed to character distortion or text-recognition failures. While standard textbooks predominantly feature canonically oriented diagrams, human cognition easily achieves viewpoint invariance via mental rotation without collapsing deductive logic; moreover, geometric theorems are fundamentally governed by topological connectivity rather than absolute canvas orientations.

The models' severe failure demonstrates that they do not execute rotation-invariant geometric deduction. Instead, they rely on rigid, fixed-orientation template matching where the relative spatial coordinates between lines and their upright text anchors are entirely re-shuffled, rendering their learned layout shortcuts completely useless under structural drift.

Referentiality: Symbolic Identities Are Fluid. Altering vertex identifiers through strict bijective permutations (e.g., mapping $\triangle ABC$ to $\triangle XYZ$)

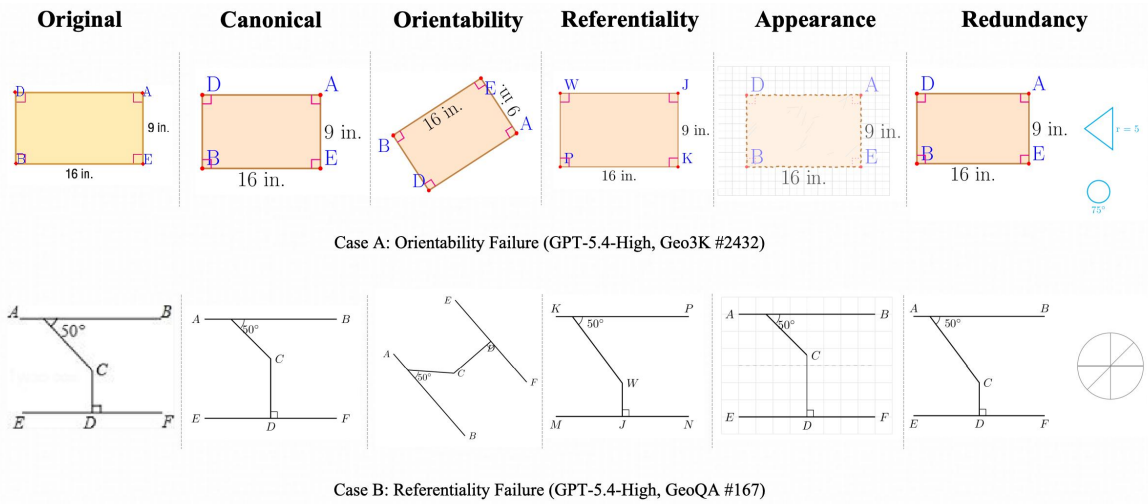

Figure 2: Illustrating the figures of representative failure cases.

breaks the high-frequency visual-textual token co-occurrences often memorized from web-scale pre-training data. Under this dimension, we observe a fascinating performance bifurcation. Advanced multi-modal models show remarkable resilience; Advanced multimodal models show strong resilience: GPT-5.4-High, Gemini-3.1-Pro, Qwen3-VL-Plus, and Claude-4.6-Opus all change by less than one point under Referentiality in Table 2.

This unexpected robustness demonstrates that top-tier closed-source models possess exceptional cross-modal grounding mechanisms. Because the logical deductive chain remains perfectly intact, their advanced vision-language alignment layers can seamlessly bypass rote textual mnemonics and perform dynamic contextual binding between the new abstract character tokens and their visual coordinate anchors.

By contrast, several open-source models remain more sensitive to relabeling, especially LLaVA-OneVision-8B and G-LLaVA-13B, which drop by 5.3 and 6.6 points respectively. This pattern suggests that dynamic symbol binding remains unevenly developed across model families.

Appearance: Stylistic Features Are Distracting. Modifications to low-level stylistic rendering—such as injecting background coordinate grids, altering line styles to dashed fragments, or applying sketch-like strokes—induce minor regressions in elite architectures but severely penalize open-source models. GPT-5.4-High and Gemini-3.1-Pro show moderate regressions of 2.1 and 1.7 points, while Qwen3-VL-Plus drops by 1.8 points.

The degradation is particularly acute for specialized open-source models; The degradation is particularly clear for MathCoder-VL-8B, which suffers the largest Appearance drop in Table 2 ($35.6\% \rightarrow 28.5\%$). From a Appearance grouping perspective, these stylistic shifts do not alter a single pixel of the underlying topological skeleton.

Top-tier models successfully demonstrate stylistic invariance by abstracting the high-level geometry away from surface-level textures. However, weaker models suffer from catastrophic attention distraction; they mistake non-semantic visual variables (like dashed auxiliary styling or background grid lines) for primary topological boundaries, which corrupts their downstream token sequence alignment and derails the geometric proof logic.

Redundancy: Visual Clutter Is Mostly Neutral. Counter-intuitively, injecting non-essential visual clutter—such as redundant tangent lines, unlinked circles, or bounding boxes placed into vacant canvas spaces—does not impose a uniform cognitive overload. Instead, it produces mixed but generally small effects across evaluated configurations. GPT-5.4-High and Gemini-3.1-Pro improve slightly under Redundancy, while InternVL-3.5-38B, LLaVA-OneVision-8B, and MathCoder-VL-8B show small declines.

Because these injected primitives are mathematically unconstrained and entirely decoupled from the core logical paths required for the solution, they do not alter the true region of interest. The empirical gains suggest that the addition of complex, irrelevant visual geometry acts as a form of structural regularization.

This mixed pattern suggests that redundant primitives mainly test attentional selectivity: some mod-

els ignore them effectively, whereas others experience mild interference.

4.3 Case Analysis

To unearth the underlying cognitive bottlenecks of multimodal geometric reasoning, we trace specific execution pipelines across individual test instances (visualized compactly in Figure 2), and isolate two foundational failure profiles that define current architectural limitations:

- **Coordinate Overfitting under Orientability:** In Case A (#2432), GPT-5.4-High fails a trivial area calculation (16×9) the moment the rectangle undergoes a rigid isometry rotation. Since the topological connectivity and side lengths are perfectly preserved, this failure diagnostic reveals that the model's visual tokenization lacks canonical viewpoint invariance. It cannot abstract geometric structures independent of absolute horizontal-vertical canvas coordinates.
- **Binding Disconnection under Referentiality:** In Case B (#167), GPT-5.4-High correctly resolves the geometric logic under standard labels (A, B, C) but suffers a catastrophic reasoning collapse under dynamic bijective remapping (K, P, W). The failure is likely associated with the *cross-modal binding layer*. The model fails to anchor the new text entities to the spatial boundaries it previously extracted, suggesting reliance on memorized textbook mnemonic co-occurrences.

4.4 Perturbation-guided Finetuning

The fine-grained diagnostic capacity of GEOPERTURB naturally motivates a critical downstream question: if layout perturbations expose highly decoupled structural vulnerabilities, can targeted, perturbation-augmented supervision actively mitigate these systemic failures? To investigate this intervention potential, we conduct a controlled sample-efficient finetuning study on the open-source `Qwen3-VL-8B` baseline. Specifically, we inject a highly restricted, low-volume partition of our programmatic `Train` configuration, augmented with our multi-dimensional layout transformations, into an adaptation pipeline. We evaluate the model using the **SFT-CoT** paradigm, which mandates the explicit generation of step-by-step symbolic and geometric derivations before outputting the final answer token.

This finetuning evaluation uses a held-out subset and a CoT-formatted evaluation prompt, so the base row is not intended to exactly reproduce the aggregate zero-shot values in Table 2.

Table 3: Perturbation-guided finetuning results on the held-out SFT evaluation subset under the CoT evaluation prompt.

Per-family accuracy				
Model	Orient.	Refer.	Appear.	Redund.
Qwen3-VL-8B	49.6	52.2	49.6	54.6
+ SFT-CoT	**57.3**	**62.1**	**61.1**	**64.7**

Table 3 reports the per-family accuracy breakdowns post-intervention. The empirical results yield an exceptional realization regarding sample efficiency: exposing the vision-language architecture to a small-scale dose of perturbation-guided supervision triggers an immediate, substantial performance surge across all evaluation frontiers. SFT-CoT comprehensively outpaces the base model across all four isolated layout families. This optimization demonstrates that the severe visual-layout fragility identified in current LMMs is not an unresolvable flaw inherent to the transformer architecture. Instead, it behaves as a latent representation misalignment that can be effectively regularized.

5 Conclusion

In this work, we have challenged the prevailing evaluation paradigms of LMMs in geometric reasoning by introducing GEOPERTURB, a programmatic framework that decouples deep logical invariants from superficial layout templates. Through the dual leverage of standardized TikZ vector reconstructions and four isolated layout perturbations, we have executed a comprehensive diagnostic audit across leading proprietary flagships and open-source models. Our empirical findings reveal that many current models remain sensitive to rigid, fixed-orientation textbook layout patterns, although the degree of vulnerability varies substantially across architectures. By shifting benchmark evaluation from static data scraping to dynamic structural perturbations, GEOPERTURB offers the community a rigorous diagnostic lens to better separate invariant spatial-logical reasoning from rendering- or template-sensitive behavior.

Limitations

While GEOPERTURB establishes a mathematically rigorous pipeline for diagnosing layout vulnerabilities, it has several limitations. Chief among these is the architectural constraints of evaluating

MLLMs as end-to-end black boxes; our current benchmarking framework can quantitatively pinpoint substantial performance degradation under spatial or symbolic drift, yet it cannot cleanly isolate whether the failure originates inside the vision encoder's micro-visual patch tokenization layers or within the large language model's downstream cross-modal semantic binding layers.

Furthermore, our generative perturbation engine is fundamentally optimized for 2D Euclidean geometry, mirroring the flat diagram distributions. Consequently, the current pipeline bypasses the higher-order topological invariants required to navigate complex 3D geometric manifolds. Bridging this dimensionality gap by extending programmatic layout perturbations into multi-dimensional and non-static geometric spaces remains a critical frontier for expanding the diagnostic boundary of multimodal spatial intelligence.

References

Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Didi Zhu, and 1 others. 2025. Llava-onevision-1.5: Fully open framework for democratized multimodal training. *arXiv preprint arXiv:2509.23661*.

Anthropic. 2026a. Claude opus 4.6 system card.

Anthropic. 2026b. Claude sonnet 4.6 system card.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, and 1 others. 2025. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Jonas Belouadi, Anne Lauscher, and Steffen Eger. 2024a. Automatikz: Text-guided synthesis of scientific vector graphics with tikz. In *International Conference on Learning Representations*, volume 2024, pages 55917–55943.

Jonas Belouadi, Simone P Ponzetto, and Steffen Eger. 2024b. Detikzify: Synthesizing graphics programs for scientific figures and sketches with tikz. *Advances in Neural Information Processing Systems*, 37:85074–85108.

Jiaqi Chen, Jianheng Tang, Jinghui Qin, Xiaodan Liang, Lingbo Liu, Eric Xing, and Liang Lin. 2021. Geoqa: A geometric question answering benchmark towards multimodal numerical reasoning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 513–523.

Luoxin Chen, Jinming Gu, Liankai Huang, Wenhao Huang, Zhicheng Jiang, Allan Jie, Xiaoran Jin, Xing Jin, Chenggang Li, Kaijing Ma, and 1 others. 2025. Seed-prover: Deep and broad reasoning for automated theorem proving. *arXiv preprint arXiv:2507.23726*.

Seunghyuk Cho, Zhenyue Qin, Yang Liu, Youngbin Choi, Seungbeom Lee, and Dongwoo Kim. 2026. Plane geometry problem solving with multi-modal reasoning: A survey. In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 110–131.

Xuanming Cui, Alejandro Aparcedo, Young Kyun Jang, and Ser-Nam Lim. 2024. On the robustness of large multimodal models against image adversarial attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24625–24634.

Deqing Fu, Ruohao Guo, Ghazal Khalighinejad, Ollie Liu, Bhuwan Dhingra, Dani Yogatama, Robin Jia, and Willie Neiswanger. 2024. Isobench: Benchmarking multimodal foundation models on isomorphic representations. In *First Conference on Language Modeling*.

Yumeng Fu, Jiayin Zhu, Lingling Zhang, Bo Zhao, Shaoxuan Ma, Yushun Zhang, Yanrui Wu, and Wenjun Wu. 2025. Geolaux: A benchmark for evaluating mllms' geometry performance on long-step problems requiring auxiliary lines. *arXiv preprint arXiv:2508.06226*.

Jiahui Gao, Renjie Pi, Jipeng Zhang, Jiacheng Ye, Wanjun Zhong, Yufei Wang, Lanqing Hong, Jianhua Han, Hang Xu, Zhenguo Li, and 1 others. 2025. G-llava: Solving geometric problem with multi-modal large language model. In *International Conference on Learning Representations*, pages 3490–3511.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Google DeepMind. 2026. Gemini 3.1 pro model card.

Christian Greisinger and Steffen Eger. 2026. Tikzilla: Scaling text-to-tikz with high-quality data and reinforcement learning. In *The Fourteenth International Conference on Learning Representations*.

Dan Hendrycks and Thomas Dietterich. 2019. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*.

Kurt Koffka. 2013. *Principles of Gestalt psychology*. routledge.

Yangyi Li, Chenxu Zhao, and Mengdi Huai. 2026. Quantifying and understanding uncertainty in large reasoning models. *arXiv preprint arXiv:2604.13395*.

Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. 2021. Inter-gps: Interpretable geometry problem solving

with formal language and symbolic reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6774–6786.

OpenAI. 2024. Gpt-4o mini: Advancing cost-efficient intelligence.

OpenAI. 2026. Introducing gpt-5.4.

Bowen Ping, Minnan Luo, Zhuohang Dang, Chenxi Wang, and Chengyou Jia. 2026. AutoGPS: Automated geometry problem solving via multimodal formalization and deductive reasoning. In *The Fourteenth International Conference on Learning Representations*.

Xiangyu Qi, Kaixuan Huang, Ashwinee Panda, Peter Henderson, Mengdi Wang, and Prateek Mittal. 2024. Visual adversarial examples jailbreak aligned large language models. In *Proceedings of the AAAI conference on artificial intelligence*, pages 21527–21536.

Jielin Qiu, Yi Zhu, Xingjian Shi, Zhiqiang Tang, Ding Zhao, Bo Li, and Mu Li. 2022. Benchmarking robustness under distribution shift of multimodal image-text models. In *NeurIPS 2022 Workshop on Distribution Shifts: Connecting Methods and Applications*.

Itsumi Saito, Haruto Yoshida, and Keisuke Sakaguchi. 2025. Sketch2diagram: Generating vector diagrams from hand-drawn sketches. In *International Conference on Learning Representations*, volume 2025, pages 14332–14354.

Nannan Shi, Chuanyu Qin, Shipeng Song, and Man Luo. 2025. Geothought: A dataset for enhancing mathematical geometry reasoning in vision-language models. *arXiv preprint arXiv:2510.21881*.

Jiayin Sun, Caixia Sun, Boyu Yang, Hailin Li, Xiao Chen, Yi Zhang, Errui Ding, Liang Li, Chao Deng, and Junlan Feng. 2026. Geotikzbridge: Advancing multimodal code generation for geometric perception and reasoning. *arXiv preprint arXiv:2603.22687*.

Boxin Wang, Chejian Xu, Shuohang Wang, Zhe Gan, Yu Cheng, Jianfeng Gao, Ahmed Hassan Awadallah, and Bo Li. 2021. Adversarial glue: A multi-task benchmark for robustness evaluation of language models. *arXiv preprint arXiv:2111.02840*.

Ke Wang, Junting Pan, Linda Wei, Aojun Zhou, Weikang Shi, Zimu Lu, Han Xiao, Yunqiao Yang, Houxing Ren, Mingjie Zhan, and Hongsheng Li. 2025a. Mathcoder-vl: Bridging vision and code for enhanced multimodal mathematical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 2505–2534.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, and 1 others. 2025b. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*.

Xiaojing Wei, Ting Zhang, Wei He, Jingdong Wang, and Hua Huang. 2025. Geoloom: High-quality geometric diagram generation from textual input. *arXiv preprint arXiv:2512.08180*.

Chen Wu and Chao Sun. 2025. Structured reasoning for geometry problems: A chain-of-knowledge approach with structurally retrieved examples. In *2025 International Conference on Intelligent Education and Intelligent Research (IEIR)*, pages 1–9. IEEE.

Shihao Xu, Yiyang Luo, and Wei Shi. 2024. Geo-llava: A large multi-modal model for solving geometry math problems with meta in-context learning. In *Proceedings of the 2nd Workshop on Large Generative Models Meet Multimodal Applications*, pages 11–15.

Liang Yan, Xu Jiang, Jian Ma, Yuhang Liu, Tian Bian, Qichao Wang, Abhishek Basu, Yu Rong, Tingyang Xu, Pengcheng Wu, and 1 others. 2025. A comprehensive survey of multimodal llms for scientific discovery. In *1st Workshop on VLM4RWD@ NeurIPS 2025*.

Jiacheng Yang, Anqi Chen, Yunkai Dang, Qi Fan, Cong Wang, Wenbin Li, Feng Miao, and Yang Gao. 2026a. Annotation-free visual reasoning for high-resolution large multimodal models via reinforcement learning. *arXiv preprint arXiv:2602.23615*.

Tianyu Yang, Sihong Wu, Yilun Zhao, Zhenwen Liang, Lisen Dai, Chen Zhao, Minhao Cheng, Arman Cohan, and Xiangliang Zhang. 2026b. A survey of multimodal mathematical reasoning: From perception, alignment to reasoning. *arXiv preprint arXiv:2603.08291*.

Yufei Zhan, Hongyin Zhao, Yousong Zhu, Shurong Zheng, Fan Yang, Ming Tang, and Jinqiao Wang. 2025. Understand, think, and answer: Advancing visual reasoning with large multimodal models. *arXiv preprint arXiv:2505.20753*.

Chengrui Zhang, Maizhen Ning, Tianyi Liu, Zihao Zhou, Jie Sun, Qiufeng Wang, and Kaizhu Huang. 2025. Geosdf: Plane geometry diagram synthesis via signed distance field. *arXiv preprint arXiv:2506.13492*.

Dachuan Zhao, Weiyue Li, Zhenda Shen, Yushu Qiu, Bowen Xu, Haoyu Chen, and Yongchao Chen. 2025a. Bias is a subspace, not a coordinate: A geometric rethinking of post-hoc debiasing in vision-language models. *arXiv preprint arXiv:2511.18123*.

Yurui Zhao, Xiang Wang, Jiahong Liu, Irwin King, and Zhitao Huang. 2025b. Towards geometry problem solving in the large model era: A survey. *arXiv preprint arXiv:2506.02690*.

Hao Zhong, Muzhi Zhu, Zongze Du, Zheng Huang, Canyu Zhao, Mingyu Liu, Wen Wang, Hao Chen, and Chunhua Shen. 2026. Omni-r1: Reinforcement learning for omnimodal reasoning via two-system collaboration. *Advances in Neural Information Processing Systems*, 38:45266–45297.

Kaijie Zhu, Qinlin Zhao, Hao Chen, Jindong Wang, and Xing Xie. 2024. Promptbench: A unified library for evaluation of large language models. *Journal of Machine Learning Research*, 25(254):1–22.

Zhihao Zhu, Jiafeng Liang, Shixin Jiang, Jinlan Fu, Ming Liu, Guanglu Sun, See-Kiong Ng, and Bing Qin. 2026. Analyzing reasoning consistency in large multimodal models under cross-modal conflicts. *arXiv preprint arXiv:2601.04073*.

Table 4: Coverage funnel. "Matched" denotes problems retained for paired original–canonical and canonical–perturbation analyses.

Subset	Candidate	Reconstructed	Coverage	Matched
GeoQA-test	49	47	95.9%	47
GeoQA-train	164	155	94.5%	155
Geo3K-test	51	49	96.1%	49
Geo3K-train	199	192	96.5%	192
Pure	90	88	97.8%	88
Total	553	531	96.0%	531

Table 5: Manual reconstruction corrections by subset. These are diagrams that failed the initial automatic reconstruction and were corrected before inclusion.

Subset	Corrected diagrams
GeoQA-test	12
GeoQA-train	9
Geo3K-test	7
Geo3K-train	25
Pure	20
Total	73

A Dataset Construction and Validation

This appendix reports construction checks, diagnostic summaries, and reproducibility details needed to interpret the main results. We avoid duplicating main-table numbers except where an additional table clarifies a matched subset, a conditional diagnostic, or a reproducibility-sensitive detail.

A.1 Coverage Funnel

Table 4 summarizes the construction funnel. We report reconstruction coverage explicitly because TikZ reconstruction is treated as a measured representation intervention rather than hidden preprocessing.

A.2 Human Audit and Manual Correction

Each included suite is inspected before evaluation. For canonical reconstructions, annotators verify geometric primitives, spatial relations, label placement, symbolic references, and answer consistency. For perturbation variants, annotators additionally verify that the intended transformation is applied and that the answer is preserved, except for Referentiality, where synchronized textual labels and symbolic answers are checked after remapping.

We do not discard difficult diagrams merely because the first automatic reconstruction fails. Instead, failed reconstructions are corrected through a human-in-the-loop TikZ editing pass and then recompiled. Table 5 reports the number of diagrams requiring manual correction. After this correction-and-audit pass, all 531 included reconstructions are eligible for downstream paired evaluation.

A.3 Scoring Validation

Model outputs are scored by deterministic answer extraction whenever possible. The fallback judge is used only when deterministic parsing cannot decide mathematical equivalence. To validate this hybrid scorer, we conduct a stratified spot-check of 200 evaluation records across models, datasets, and perturbation families. A human annotator independently re-scores each record using the problem, gold answer, model response, and automatic verdict. Human–automatic agreement is 96.5% with Cohen's $\kappa = 0.93$. Most disagreements involve equivalent symbolic forms or borderline numeric tolerance, such as $\sqrt{3}$ versus 1.732.

B Supplementary Robustness Summary

The main paper reports original, canonical, joint perturbation, and family-level accuracy. This section adds one compact delta table to make the direction of each perturbation family explicit. The deltas in Table 6 use the same matched aggregate as Table 2. Since the main table rounds accuracies to one decimal, some deltas are computed before rounding.

B.1 Dimension-Level Accuracy Deltas

For each perturbation family f, we report

$$\Delta_f = A_f - A_{\text{canonical}}.$$

Negative values indicate accuracy loss relative to the canonical reconstruction, while positive values indicate accuracy gain.

C Referential Grounding Diagnostics

Referentiality is implemented through synchronized label remapping. In this appendix we refer to this implementation as SymbolMapping. It preserves the underlying geometry while changing the label-to-entity bindings, and therefore provides a direct probe of symbolic grounding.

C.1 Conditional Symbol Drop

For Referentiality, we report Conditional Symbol Drop:

$$\text{CSD} = P(\hat{y}^{\text{SM}} \neq y^{\text{SM}} \mid \hat{y}^{\text{rec}} = y).$$

CSD asks: among instances that a model solves under the canonical reconstruction, how often does

Table 6: Per-dimension accuracy delta relative to Canonical (%). Negative means the perturbation hurts; positive means it helps.

Model	Orientability	Referentiality	Appearance	Redundancy
Proprietary Models				
GPT-5.4-High	-4.3	-0.8	-2.1	$+1.8$
Gemini-3.1-Pro	-5.3	-0.2	-1.7	$+1.0$
Qwen3-VL-Plus	-6.4	-0.2	-1.8	$+0.1$
Claude-4.6-Opus	-6.8	-0.5	$+0.3$	-1.0
Claude-4.6-Sonnet	-6.4	-2.3	$+0.6$	-0.1
GPT-4o-mini	-1.1	-0.8	-0.8	-0.3
Open-source Models				
Qwen3-VL-8B	-4.1	-2.3	-4.6	$+0.8$
InternVL-3.5-38B	-5.4	-4.4	-2.1	-2.2
LLaVA-OneVision-8B	-4.0	-5.3	-1.8	-1.2
MathCoder-VL-8B	-2.3	-2.4	-7.1	-1.8
G-LLaVA-13B	$+0.9$	-6.6	-0.7	$+1.2$

the same model fail after consistent symbolic relabeling? This conditioning reduces confounding from general problem difficulty and focuses the diagnostic on referential grounding.

Table 7 reports CSD for the proprietary models. The denominator n_{SM} is not the full split size; it is the number of Referentiality-valid instances that were solved under canonical reconstruction and therefore enter the conditional computation. $A_{\mathrm{SM|rec}}$ is Referentiality accuracy restricted to this reconstructed-correct subset, so $\mathrm{CSD} = 100 - A_{\mathrm{SM|rec}}$ up to rounding.

These conditional quantities answer a stricter question than the marginal Referentiality deltas in Table 2. A model can have a small marginal Referentiality drop while still exhibiting non-trivial CSD, because some originally canonical-wrong instances may become correct after relabeling.

C.2 Mechanism Audit: Perception vs. Binding

To separate label perception from symbolic binding, we conduct a targeted audit on Referentiality failures. A failure case is a model–problem pair where the model answers the canonical reconstruction correctly but answers the Referentiality version incorrectly. We sample 20 such pairs, covering 16 unique problems and five proprietary models.

Label-reading probe. For each failure case, the model is shown the Referentiality diagram and asked to list all visible point labels. The reported labels are compared against labels extracted from the perturbed TikZ code. Among 18 valid responses, mean recall is 91.4%. This indicates that many audited Referentiality failures are not simply caused by missing or unreadable labels.

Oracle mapping recovery. For the same 20 failure cases, we provide an explicit original-to-current label mapping table before solving. Only 5 of 20 cases recover the correct answer, as shown in Table 8. Thus, even when the remapping is made explicit, most failures persist.

Interpretation. The combination of high label recall and low oracle recovery suggests that many Referentiality failures occur after local perception: the model reads the new labels but does not reliably propagate updated symbol-to-entity bindings through multi-step reasoning. Because the audit is targeted and small, we use it as a mechanism probe rather than as a standalone population estimate.

D Split Interpretation and Template-Reliance Screen

Train–test comparisons cannot prove data contamination, because train examples can differ in difficulty, rendering style, and diagram distribution. We therefore treat train–test perturbation asymmetry only as a behavioral screen: a larger train advantage under answer-preserving perturbations is consistent with presentation familiarity, but it is not forensic evidence of pretraining overlap.

The Pure split is used as a leakage-mitigated and distribution-shifted evaluation anchor. It should not be interpreted as absolutely absent from proprietary pretraining corpora. Accordingly, all split-level interpretations in the main paper are phrased as robustness diagnostics rather than as direct claims about training-data membership.

Table 7: Referentiality grounding diagnostic. $A_{\mathrm{SM|rec}}$ is Referentiality accuracy among reconstructed-correct instances; CSD is the corresponding conditional failure rate. All values are percentages.

Dataset	Model	n_{SM}	$A_{\mathrm{SM\|rec}}$	CSD
Geo3K-test	GPT-5.4-High	31	90.3	9.7
	Gemini-3.1-Pro	45	97.8	2.2
	Qwen3-VL-Plus	23	95.7	4.3
	Claude-4.6-Opus	34	91.2	8.8
	Claude-4.6-Sonnet	43	97.7	2.3
	GPT-4o-mini	21	71.4	28.6
GeoQA-test	GPT-5.4-High	33	78.8	21.2
	Gemini-3.1-Pro	46	84.8	15.2
	Qwen3-VL-Plus	43	90.7	9.3
	Claude-4.6-Opus	33	87.9	12.1
	Claude-4.6-Sonnet	30	90.0	10.0
	GPT-4o-mini	12	58.3	41.7
Pure	GPT-5.4-High	57	75.4	24.6
	Gemini-3.1-Pro	61	90.2	9.8
	Qwen3-VL-Plus	61	77.0	23.0
	Claude-4.6-Opus	52	73.1	26.9
	Claude-4.6-Sonnet	56	78.6	21.4
	GPT-4o-mini	25	56.0	44.0
Geo3K-train	GPT-5.4-High	168	93.5	6.5
	Gemini-3.1-Pro	176	96.6	3.4
	Qwen3-VL-Plus	169	92.3	7.7
	Claude-4.6-Opus	172	96.5	3.5
	Claude-4.6-Sonnet	160	90.0	10.0
	GPT-4o-mini	93	75.3	24.7
GeoQA-train	GPT-5.4-High	135	93.3	6.7
	Gemini-3.1-Pro	146	95.9	4.1
	Qwen3-VL-Plus	134	88.1	11.9
	Claude-4.6-Opus	135	89.6	10.4
	Claude-4.6-Sonnet	129	87.6	12.4
	GPT-4o-mini	50	66.0	34.0

Table 8: Oracle mapping recovery in the Referentiality mechanism audit.

Model	Recovered	Total
Claude-4.6-Opus	1	5
Claude-4.6-Sonnet	0	2
GPT-4o-mini	1	4
GPT-5.4-High	1	5
Qwen3-VL-Plus	2	4
Total	5	20

E Statistical Calibration

All paired comparisons are computed over matched problem suites. For reconstruction effects, comparisons are paired by problem between the original and canonical views. For perturbation-family effects, comparisons are paired between the canonical view and the corresponding family-specific perturbation. For the joint Perturbed metric in Table 1, correctness is evaluated at the problem-suite level: an instance is counted as correct only when all four perturbation families are solved.

Since the paper contains many model–dataset cells, individual cells are interpreted descriptively unless supported by large effect sizes and consistent cross-family patterns. The main conclusions rely on repeated patterns across model families and perturbation dimensions rather than on a single isolated cell.

F Qualitative Diagnostics

We include representative cases that illustrate separable failure modes. These examples are diagnostic rather than exhaustive.

Referentiality failure. In GeoQA #167, GPT-5.4-High solves the canonical reconstruction but fails after synchronized label remapping. The geometry is unchanged; the failure is triggered by changing the symbolic identities attached to the same visual entities. This example illustrates why Referentiality is a direct test of referential grounding.

Orientability failure. In Geo3K #2432, GPT-5.4-High solves the canonical rectangle-area prob-

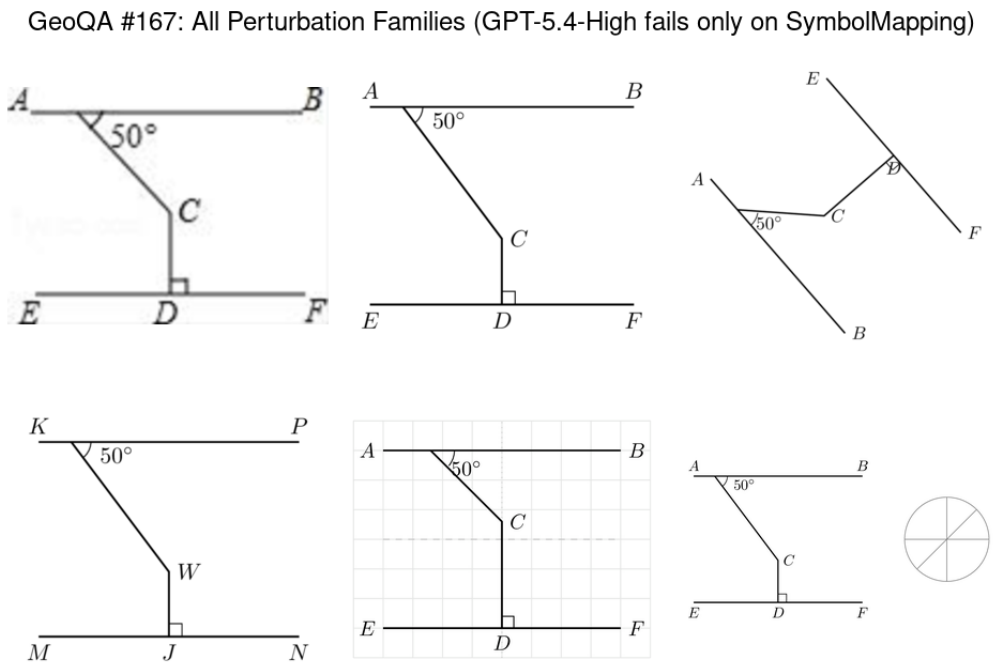

Figure 3: GeoQA #167: representative Referentiality failure under synchronized label remapping.

lem but fails after a rigid rotation that preserves side lengths, right angles, and the gold answer. This isolates viewpoint sensitivity from label OCR and appearance changes.

Robust triplet. In GeoQA #0, GPT-5.4-High solves the original, canonical, and perturbed views using the same parallel-line angle relation. This case shows that GEOPERTURB does not merely manufacture failures: robust behavior is achievable when the model maintains stable visual-symbolic bindings.

Pure split algebraic geometry. The Pure split also includes algebraic geometry instances rather than only angle-chasing problems. Figure 4 illustrates a tile-area problem where perturbations preserve the geometric layout while exposing model-specific reasoning instability.

G Perturbation-Guided Finetuning Details

The finetuning experiment in Section 4.4 is intended as a targeted intervention study rather than as a full leaderboard replacement. The base and SFT-CoT rows in Table 3 are evaluated on a held-out SFT evaluation subset under a CoT-formatted evaluation prompt. Therefore, the base row in that table is not expected to exactly reproduce the aggregate zero-shot values in Table 2.

The purpose of the intervention is to test whether exposure to perturbation-augmented supervision can reduce family-specific fragility. The results should be interpreted as evidence that some structural vulnerabilities are at least partially mitigable under targeted supervision, not as proof that the underlying reasoning problem is fully solved.

H Reproducibility Details

H.1 Model Identifiers and Decoding

Table 9 lists the display names and identifiers used in the evaluation. All models are evaluated with deterministic decoding, temperature 0, and a maximum output budget of 8192 tokens. For local open-source models, inference is run through the corresponding local endpoint. For provider models, evaluation is run through the provider API.

H.2 Prompt Templates

Reconstruction prompt.

```
You are a Geometry-to-TikZ Transpiler.
Convert the input geometry image into
precise TikZ code.  Output only the
TikZ picture body.  Do not include a
document preamble, markdown, comments,
or explanatory text. Preserve geometric
primitives, right-angle markers, angle
arcs, labels, point markers, and
numerical annotations. Ensure labels do
not overlap whenever possible.
```

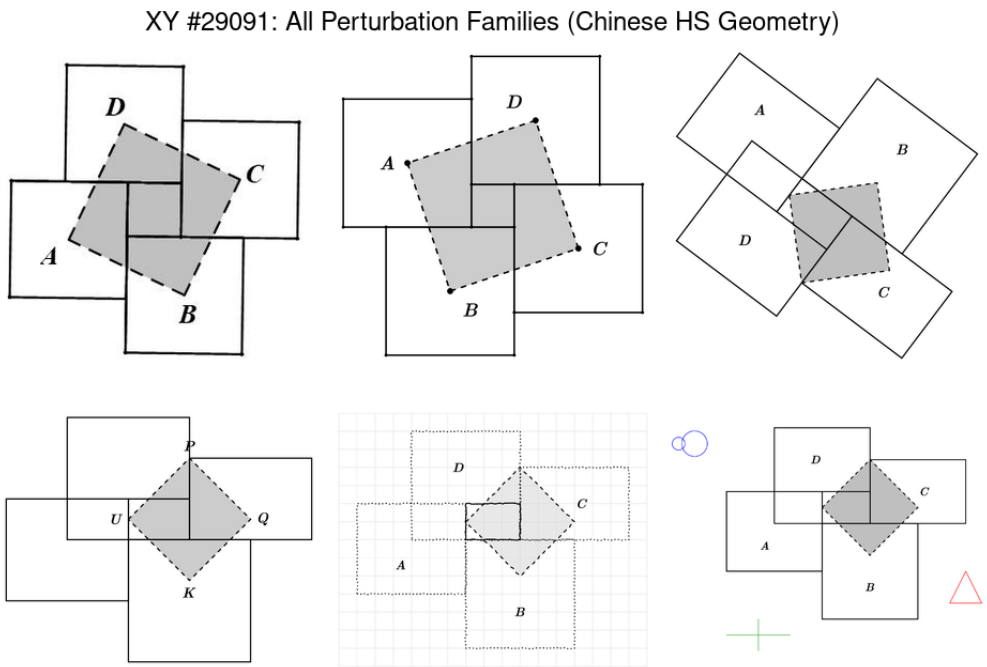

Figure 4: Pure split example #29091: algebraic geometry problem with four perturbation families.

Table 9: Model identifiers, providers, and evaluation configurations.

Display name	API / checkpoint identifier	Provider	Type	Datasets
Proprietary Models				
GPT-5.4-High	`gpt-5.4` with `reasoning_effort=high`	OpenAI	API	all
GPT-4o-mini	`gpt-4o-mini`	OpenAI	API	all
Claude-4.6-Opus	`claude-opus-4-6-thinking`	Anthropic	API	all
Claude-4.6-Sonnet	`claude-sonnet-4-6`	Anthropic	API	all
Gemini-3.1-Pro	`gemini-3.1-pro-preview`	Google	API	all
Qwen3-VL-Plus	`qwen3-vl-plus`	Alibaba	API	all
Open-source Models				
Qwen3-VL-8B	`Qwen/Qwen3-VL-8B-Instruct`	Alibaba / local checkpoint	local	all
InternVL-3.5-38B	`OpenGVLab/InternVL3-38B`	Shanghai AI Lab / local checkpoint	local	all
LLaVA-OneVision-8B	`lmms-lab/llava-onevision-qwen2-7b-ov`	LMMs-Lab / local checkpoint	local	all
MathCoder-VL-8B	`MathCoder-VL-8B`	local checkpoint	local	all
G-LLaVA-13B	`G-LLaVA-13B`	local checkpoint	local	all

Reconstruction refinement prompt.

```
Compare the original image with the rendered TikZ image. Check geometric primitives, right-angle markers, angle arcs, labels, point markers, segment labels, and spatial relations. If the render matches the original sufficiently for solving the problem, output only <PERFECT>. Otherwise, output corrected TikZ code only.
```

Perturbation prompt.

```
You are creating a structure-preserving geometry perturbation. Target family: {family}. Preserve all lengths, angles, incidences, and answer-critical relations unless the family explicitly requires synchronized symbolic relabeling. The gold answer must remain unchanged except for Referentiality, where labels in the question, diagram, options, and symbolic answer must be updated consistently. Output only the perturbed TikZ code.
```

Zero-shot evaluation prompt.

```
Solve the geometry problem using the provided diagram and text. Return the final answer in the following JSON format: {"answer": "<final answer>"}. The answer should be a single option, number, expression, or symbolic value whenever possible.
```

SFT-CoT evaluation prompt.

```
Solve the geometry problem using the provided diagram and text. Provide concise step-by-step geometric
```

reasoning, then return the final answer in the following JSON format: {“answer”: “<final answer>”}.

Fallback judging prompt.

You are grading a geometry answer. Given the problem, gold answer, and model response, decide whether the response contains a mathematically equivalent final answer. Ignore harmless formatting differences. Output only CORRECT or INCORRECT.